Variational Inference (Contd)

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Feb 13, 2019

Announcements

- Mid-sem exam on Monday, Feb 18, 8:00am-10:00am (L-19, ERES)
- Syllabus up to today's lecture (but mostly the basics of VI)
- Questions will be a mix of MCQ, fill-in-the-blanks, short answer, and not-so-short answer
- Answer must be written on the question paper itself in provided space
- Advised to use pencil and eraser (but write prominently)
- The exam will be closed book and closed notes/slides
- Necessary formulae/results etc will be provided in the question paper itself
- A revision-cum-QA session on Friday (or Saturday?) 6:30pm in KD-101

Recap: VI and Mean-Field VI

- Approximate an intractable posterior $p(\mathbf{Z}|\mathbf{X})$ by another distribution $q(\mathbf{Z}|\phi)$ by solving

$$\phi^* = \arg\min_{\phi} \mathrm{KL}[q_\phi(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})] \qquad \text{or equivalently} \qquad q^*(\mathbf{Z}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}[q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})]$$

- Equivalent to finding q that maximizes the Evidence Lower Bound (ELBO)

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Can further simplify using a mean-field assumption on q: $q(\mathbf{Z}|\phi) = \prod_{j=1}^{M} q(\mathbf{Z}_j|\phi_j)$
- For the optimal q_j, $\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, and thus

$$q_j^*(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) d\mathbf{Z}_j} \propto \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) \qquad \forall j$$

- Mean-field VI updates the q_j's in a cyclic manner, like ALT-OPT, Gibbs sampling, etc

Mean-Field VI: A Very Simple Example

- Consider data $\mathbf{X} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ with mean μ, precision τ
- Assume the following normal-gamma prior on μ and τ

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) \quad p(\tau) = \text{Gamma}(\tau|a_0, b_0)$$

- Note: Here posterior is straightforward (normal-gamma due to the jointly conjugate prior)
- Let's try mean-field VI nevertheless to illustrate the idea
- With mean-field assumption on the variational posterior $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \end{aligned}$$

- In this example, the log-joint $\log p(\mathbf{X}, \mu, \tau) = \log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$. Therefore

$$\begin{aligned} \log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau)] + \text{const} \qquad \text{(only keeping terms that involve } \mu\text{)} \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)] + \text{const} \end{aligned}$$

Mean-Field VI: A Very Simple Example (Contd)

- Substituting the expressions $p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} p(x_n|\mu,\tau)$ and $\log p(\mu|\tau)$, we get

$$\begin{aligned}\log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu,\tau) + \log p(\mu|\tau)] + \text{const} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2}\left\{\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right\} + \text{const}\end{aligned}$$

- (Verify) The above is log of a Gaussian. Thus $q_\mu^*(\mu) = \mathcal{N}(\mu|\mu_N,\tau_N)$ with

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \quad \text{and} \quad \lambda_N = (\lambda_0 + N)\mathbb{E}_{q_\tau}[\tau]$$

- Proceeding in a similar way (verify), we can show that $q_\tau^*(\tau) = \text{Gamma}(\tau|a_N, b_N)$

$$a_N = a_0 + \frac{N+1}{2} \quad \text{and} \quad b_N = b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

- Important: Updates of $q_\mu^*(\mu)$ and $q_\tau^*(\tau)$ depend on each-other (thus requires cyclic updates)

Mean-Field VI: A Closer Look

- Since $\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z}_j, \mathbf{Z}_{-j})] + \text{const}$, we can also write

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j})] + \text{const}$$

- This is interesting: The form of optimal $q_j(\mathbf{Z}_j)$ will be the same as the conditional posterior of $\mathbf{Z}_j$
- For locally conjugate models, $p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j})$ is easy to find, and usually an exp-fam dist.

$$p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j}) = h(\mathbf{Z}_j) \exp\left[\eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j}))\right]$$

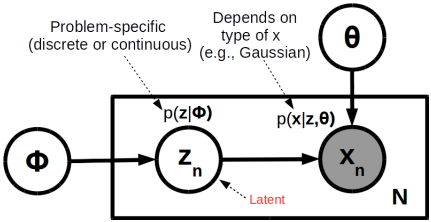
  where $\eta()$ denotes the natural params of this exp-fam distribution (would depends on $\mathbf{X}$ and $\mathbf{Z}_{-j}$)
- Using the above, we can rewrite the optimal variational distribution as follows

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}\left[\log\left(h(\mathbf{Z}_j) \exp\left[\eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j}))\right]\right)\right] + \text{const}$$

$$\implies q_j^*(\mathbf{Z}_j) \propto h(\mathbf{Z}_j) \exp\left[\mathbb{E}_{i \neq j}[\eta(\mathbf{X}, \mathbf{Z}_{-j})]^\top \mathbf{Z}_j\right] \quad \text{(verify)}$$

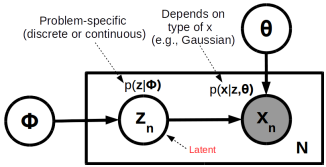
- So, in exp-fam case, getting $q_j^*(\mathbf{Z}_j)$ just requires expectation of nat. params. of cond. post. of $\mathbf{Z}_j$
- Important/useful to keep these facts in mind (will use these later)

VI for Models with Local and Global Variables



- Many LVMs consists of local variables $\mathbf{Z}$ and global variables $\boldsymbol{\beta}$ (θ, ϕ above),, e.g.,
  - GMM: $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$ are cluster ids, $\boldsymbol{\beta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \}_{k=1}^K$
  - PPCA: $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$ are latent codes, $\boldsymbol{\beta}$ are params defining the "decoder" ($\boldsymbol{z}_n$ to $\boldsymbol{x}_n$ mapping)

VI for Models with Local and Global Variables (Contd)



- Assuming independence, the joint distribution of data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and unknowns $\mathbf{Z}, \boldsymbol{\beta}$

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\beta}) p(\boldsymbol{z}_n|\boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta})$$

- Assume the joint dist. of data $\boldsymbol{x}_n$ and local var $\boldsymbol{z}_n$ is an exp-fam dist with global params $\boldsymbol{\beta}$

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\beta}) = h(\boldsymbol{x}_n, \boldsymbol{z}_n) \exp\left[\boldsymbol{\beta}^\top t(\boldsymbol{x}_n, \boldsymbol{z}_n) - A(\boldsymbol{\beta})\right]$$

- Assume a prior on global variables $\boldsymbol{\beta}$, that is conjugate to the above exp-fam dist

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = h(\boldsymbol{\beta}) \exp\left[\boldsymbol{\alpha}^\top [\boldsymbol{\beta}, -A(\boldsymbol{\beta})] - A(\boldsymbol{\alpha})\right]$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^\top$ are hyperparams of the prior $p(\boldsymbol{\beta})$ and $[\boldsymbol{\beta}, -A(\boldsymbol{\beta})]$ is the suff-stats vector

VI for Models with Local and Global Variables (Contd)

- Let's derive mean-field VI for such models
- To do so, we need the conditional posterior of each local/global variable
- Conditional posterior of global vars $\boldsymbol{\beta}$, will be in the same family as their prior $p(\boldsymbol{\beta}|\boldsymbol{\alpha})$

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{Z}) = p(\boldsymbol{\beta}|\hat{\boldsymbol{\alpha}}) \quad \text{where} \quad \hat{\boldsymbol{\alpha}} = \left[\alpha_1 + \sum_{n=1}^{N} t(\boldsymbol{x}_n, \boldsymbol{z}_n), \alpha_2 + N\right]^\top$$

- Conditional posterior of each local variable $\boldsymbol{z}_n$ will be

$$p(\boldsymbol{z}_n|\mathbf{Z}_{-n},\mathbf{X},\boldsymbol{\beta}) = p(\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{\beta}) \qquad \text{(assuming independence)}$$

- Assume the above CP to be an exp-fam dist (will usually be if $p(\boldsymbol{x}_n|\boldsymbol{z}_n)$ and $p(\boldsymbol{z}_n)$ are in exp-fam)

$$p(\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{\beta}) = h(\boldsymbol{z}_n)\exp\left[\eta(\boldsymbol{x}_n,\boldsymbol{\beta})^\top \boldsymbol{z}_n - A(\eta(\boldsymbol{x}_n,\boldsymbol{\beta}))\right]$$

- With the CPs for $\boldsymbol{\beta}$ and $\boldsymbol{z}_n$'s, deriving the mean-field VI updates for these is easy!

VI for Models with Local and Global Variables (Contd)

- Let's assume our mean-field approximation to be of the form

$$q(\boldsymbol{\beta}, \mathbf{Z}) = q(\boldsymbol{\beta}|\lambda) \prod_{n=1}^{N} q(\boldsymbol{z}_n|\phi_n)$$

- Also, here CPs are ex-fam, so optimal q's depend on expected suff-stats of CP's nat. params
- The optimal variational dist. for local vars $\boldsymbol{z}_n$ will be $q(\boldsymbol{z}_n|\phi_n)$ with

$$\phi_n = \mathbb{E}_\lambda \left[\eta(\boldsymbol{x}_n, \boldsymbol{\beta})\right] \qquad \forall n$$

- The optimal variational dist. for global vars $\boldsymbol{\beta}$ will be $q(\boldsymbol{\beta}|\lambda)$ with

$$\lambda = \left[\alpha_1 + \sum_{n=1}^{N} \mathbb{E}_{\phi_n}[t(\boldsymbol{x}_n, \boldsymbol{z}_n)], \alpha_2 + N\right]^\top$$

- The mean-field VI algo iterates b/w estimating ϕ_n's $\forall n$, and λ, until ELBO value converges
- A potential bottleneck: Updating λ requires waiting for all ϕ_n's to be updated (slow for large N)
  - But this can be handled by online VI (a.k.a. stochastic variational inference - SVI); akin to online EM
  - We will look at SVI (along with other advanced VI methods) after mid-sem

VI by Taking ELBO Gradients

- More general way of doing VI is by computing ELBO's gradient and doing gradient ascent/descent
- The gradient based approach is broadly applicable, not just for mean-field VI. Works as follows
  1. Assume $q(\mathbf{Z})$ to be from some family of distributions with variational parameters ϕ
  2. Write down the **full ELBO** expression (this will give us a function of variational params ϕ)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

  3. Compute **ELBO gradients**, i.e., $\nabla_\phi \mathcal{L}(\phi)$ and use gradient methods to find optimal ϕ

- Note: Step 2 may be simplified due to the problem structure or assumptions on the form of $q(\mathbf{Z})$
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$; conditionally independent priors simplify $\log p(\mathbf{Z})$
  - Locally-conjugate models
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$
  - Note that the last term reduces to sum of entropies of q_i's (which usually has known forms)

Posterior Predictive with VI Approximations

- Given a VI approximation of the posterior, we can use it to approximate the posterior predictive
- For example, for a K component GMM, suppose we use the following form of variational posterior

$$p(\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}_{k=1}^K) = q^*(\boldsymbol{\pi}) \prod_{k=1}^K q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$$

- The mean-field VI updates will be as follows (PRML Sec 10.2)

$$q^\star(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \qquad \alpha_k = \alpha_0 + N_k$$

$$q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)$$

$$\begin{aligned}
\beta_k &= \beta_0 + N_k \\
\mathbf{m}_k &= \frac{1}{\beta_k}(\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k) \\
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}} \\
\nu_k &= \nu_0 + N_k.
\end{aligned}$$

Posterior Predictive with VI Approximations

- Given a new observation $\hat{\boldsymbol{x}}$ and past data $\mathbf{X}$, the *true* posterior predictive for a GMM is

$$p(\widehat{\mathbf{x}}|\mathbf{X}) = \sum_{\widehat{\mathbf{z}}} \iiint p(\widehat{\mathbf{x}}|\widehat{\mathbf{z}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\widehat{\mathbf{z}}|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) \, \mathrm{d}\boldsymbol{\pi} \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Lambda}$$

$$p(\widehat{\mathbf{x}}|\mathbf{X}) = \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}\left(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) \, \mathrm{d}\boldsymbol{\pi} \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Lambda}$$

- Given the variational approx. of posterior, the posterior predictive can be approximated as

$$p(\widehat{\mathbf{x}}|\mathbf{X}) = \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}\left(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \, \mathrm{d}\boldsymbol{\pi} \, \mathrm{d}\boldsymbol{\mu}_k \, \mathrm{d}\boldsymbol{\Lambda}_k$$

$$p(\widehat{\mathbf{x}}|\mathbf{X}) = \frac{1}{\widehat{\alpha}} \sum_{k=1}^{K} \alpha_k \mathrm{St}(\widehat{\mathbf{x}}|\mathbf{m}_k, \mathbf{L}_k, \nu_k + 1 - D)$$

$$\mathbf{L}_k = \frac{(\nu_k + 1 - D)\beta_k}{(1 + \beta_k)} \mathbf{W}_k$$

Some Properties of VB

Recall that VB is equivalent to finding q by minimizing $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q||p) = \int q(\mathbf{Z}) \log \left[\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \right]$$

If the true posterior $p(\mathbf{Z}|\mathbf{X})$ is very small in some region then, to minimize $\mathrm{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

This has two key consequences for VB

- Underestimates the variances of the true posterior
- For multimodal posteriors, VB locks onto one of the modes

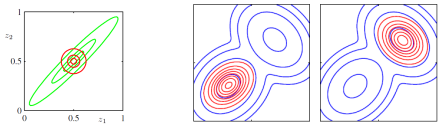
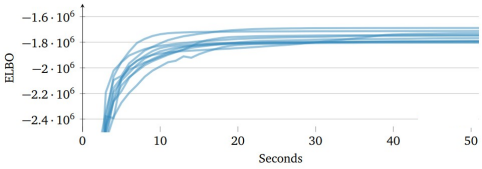


Figure: (Left) Zero-Forcing Property of VB, (Right) For multi-modal posterior, VB locks onto one of the models

Note: Some other inference methods, e.g., Expectation Propagation (EP) can avoid this behavior

VI and Convergence

- VI is guaranteed to converge but only to a local optima (just like EM)
- Therefore proper initialization is important (just like EM)



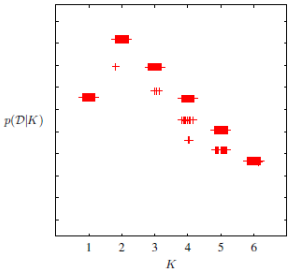
Different initializations may lead to different optima

- ELBO increases monotonically with iterations, so we can monitor the ELBO to assess convergence

ELBO for Model Selection

- Recall that ELBO is a lower bound on log of model evidence $\log p(\mathbf{X}|m)$)
- We can compute ELBO for each model m and then choose the one with largest value of ELBO
- An Example: The ELBO plot for a GMM with different K values (number of components)

Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data, showing a distinct peak at $K = 2$ components. For each value of K, the model is trained from 100 different random starts, and the results shown as '+' symbols plotted with small random horizontal perturbations so that they can be distinguished. Note that some solutions find suboptimal local maxima, but that this happens infrequently.



- Note that unlike likelihood, ELBO doesn't monotonically increase with K (penalizes large K)
- Some criticism since we are using a lower-bound but works well in practice in many problems

Figure courtesy: PRML (Bishop, 2006)

VI and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VI there is no distinction between parameters Θ and latent variables $\mathbf{Z}$
- VI treats all unknowns of the model as latent variables and calls them $\mathbf{Z}$
- Since there is no notion of "parameters", VI is like EM without the "M step"
- VI can be used within an EM algorithm if the E step is intractable
  - This is known as Variational EM algorithm

VI: The Road Ahead

- Moving beyond locally conjugate models
- Moving beyond the mean-field assumption
- More scalable variational inference
- General-purpose VI (that doesn't require model-specific derivations)
  - Posing VI as a general gradient based optimization problem

$$\phi^{new} = \phi^{old} + \eta \times \nabla_{\phi} \left[\mathbb{E}_{q_\phi}[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_{q_\phi}[\log q(\mathbf{Z}|\phi)] \right]$$

  - A lot of recent research on approximating the gradient of an expectation
- We will look at these issues after mid-sem